DOCUMENT RESUME

ED 349 321 TM 018 944

AUTHOR Jaeger, Richard M.; Usher, Claire H.
TITLE Alternative Procedures for Integrating Multidimensional Evaluations of Schools: An Experimental Comparison.
PUB DATE Apr 91
NOTE 38p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 3-7, 1991).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Achievement Tests; Comparative Analysis; *Curriculum Evaluation; Educational Assessment; *Educational Quality; Elementary Secondary Education; *Evaluation Methods; *Evaluators; *Institutional Characteristics; Interrater Reliability; Standardized Tests
IDENTIFIERS *Multi Attribute Utility Theory; *Policy Capturing Method; Weighting (Statistical)

ABSTRACT

This paper reports on a study of the foundation and application of two procedures used to specify appropriate weights to be applied to components in determining the overall quality of a school. These procedures are multiattribute utility technology (MAUT) and policy capturing, and the paper presents the results of applying them, using key educational leaders in a southern state as judges. School components considered were: (1) selected core academic programs; (2) selected programs for specific subpopulations (special education and at-risk students); (3) a service delivery program; (4) selected administrative programs; and (5) several averaged student performances on standardized achievement tests. Using MAUT, judges identify attributes, determine their relative importance, estimate the value of each entity, derive an overall judgment for each, and investigate the consistency and sensitivity of judgments produced. Policy capture describes the unique information processing strategies of individual raters statistically. Judgment data were collected from 29 educational leaders through MAUT and policy capturing. Weights obtained from both procedures indicate that assigning equal weights to the attributes would misrepresent the values judges ascribed to these attributes as contributors to the overall quality of schools. Each procedure produced distributions of median weights that varied substantially from a uniform distribution, and analysis of the differences between them indicated that, if consistency of judgment is the criterion for selection of an attribute weighting method, policy capture is the method of choice. Two tables and four figures, which are replicated in an attached handout, present study data. There is a 61-item list of references. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

RICHARD M. JAEGER

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED349321

Alternative Procedures for Integrating Multidimensional Evaluations of Schools: An Experimental Comparison

Richard M. Jaeger
Claire H. Usher
Center for Educational Research and Evaluation
University of North Carolina at Greensboro

Presented at the 1991 Annual Meeting of the American Educational Research Association, Chicago, April 3-7

TM018944

BEST COPY AVAILABLE

The problem addressed in this paper arose in a decision context that required determination of the overall quality of public elementary and secondary schools on the basis of information concerning the quality of various constituent attributes of those schools (which we have termed components). The components to be considered included selected core academic programs (reading/language arts programs and mathematics programs), selected programs for specific subpopulations of students (special education programs and programs for "at-risk" students), a service-delivery program (guidance and counseling), selected administrative programs (personnel evaluation and development programs), and several averaged student performances on standardized achievement tests (reading/language arts tests and mathematics tests). The problem was to determine the relative weights to be applied to these components in determining the overall quality of a school.

The solution to this problem requires that human judgment be used to determine appropriate weights. We contend that the procedures used to determine weights should be explicit, public, and replicable. In addition, such procedures should incorporate the values held by a sample of appropriate judges; they should not reflect unduly the values held by any individual (Edwards, 1977; p. 326). In this paper we describe the foundation and application of two procedures that have been used to specify appropriate weights: multiattribute utility technology and policy capturing. We also present the results of applying these procedures, using as judges a group of key educational leaders in a southeastern state.

Background

The problem of eliciting and combining judgments of the quality or desirability of components (attributes) for the purpose of forming judgments of the quality or desirability of an entity has received substantial attention in social psychology, industrial and organizational psychology, and the field of operations research (Dyer & Sarin, 1979; Edwards & Newman, 1982; Eils & John, 1980; Einhorn

& McCoach, 1977; Gardiner & Edwards, 1975; Keeney, 1977, 1982; Pitz, 1980; Wiener, Frank, & Libros, 1988). Common entities that have been used as the objects of judgment include persons (e.g., persons applying for jobs or persons applying for admission to an education program), social service programs (e.g., drug rehabilitation programs), and institutions (e.g., municipal government agencies). The most popular judgmental approach to evaluation of such entities carries the label "multiattribute utility theory" (Einhorn & McCoach, 1977; Yates & Jagacinski, 1979; Gardiner & Edwards, 1975) or "multiattribute utility technology," both abbreviated MAUT (Edwards & Newman, 1982). Although the approach has been widely used in the fields mentioned earlier, applications in the educational arena have been limited (one educational application of MAUT is described in Wiener, Frank, & Libros, 1988).

Multiattribute Utility Technology

A MAUT procedure consists of a well-defined sequence of steps. First, expert judges identify the attributes that will be used to evaluate the entities under consideration. For example, if the entities are job applicants, the attributes might be years of relevant education, prior work experience, prior achievements, test performances, recommendations, etc. Second, the judges determine the relative importance of each attribute to an overall judgment of the quality of the entities. In the example just cited, judges would determine how much more (or less) important an applicant's years of relevant education was than his/her prior work experience, in judging the applicant's desirability as a future employee. These kinds of relative importance judgments are made for each pair of attributes. Third, judges estimate the value of the location of each entity under consideration, on each attribute identified in Step 1. For example, they would determine the value (utility) of the education level, work experience, etc. of each job applicant being considered. This step is known as developing utility functions for the attributes. Fourth, the

judgments of relative importance of attributes produced in Step 2 are combined with the judgments of value produced in Step 3, to derive an overall judgment of value of each entity under consideration (e.g., each job applicant). The entity judged to be highest in value is the one that is typically selected. Fifth (and finally), the consistency and sensitivity of the judgments produced in Steps 2 and 3 are investigated, to determine their influence on the results of Step 4. That is, the relative importance values assigned to the attributes are varied across a limited range so as to determine the effect of such variations on the overall value of the entity being judged; i.e., the outcome of Step 4.

MAUT was developed to address decision problems in which a choice was to be made among definable alternatives, such as hiring the best of a number of applicants for a job or selecting the best of a number of potential sites for a new factory. Since the present decision problem does not involve selection among alternatives, it cannot be solved by applying MAUT in its entirety. However, the decision problem presented by the need to determine the quality of schools requires consideration of multiple attributes and specification of the relative importance of these attributes in determining the overall quality of a school. One element of the MAUT procedure is designed to address such problems explicitly.

Determining the Relative Importance of Attributes using MAUT.

Several procedures have been proposed by Edwards and Newman (1982) for determining the relative importance of attributes to an overall judgment of the quality of an entity. The one used for this study, which involves independent consideration of the relative importance of any pair of attributes, is termed "ratio weighting," and consists of the following steps:

1. The attributes are listed in rank order of their judged importance, with the least-important attribute listed first.

2. The least-important attribute is assigned an initial importance value of 100.

3. The next-least-important attribute is assigned an initial value that reflects its relative importance, compared to the least-important attribute. Thus, if the next-least-important attribute was thought to be half again as important as the least-important attribute, it would be assigned an initial value of 150, which is 50 percent larger than the value assigned to the least-important attribute (100). If the next-least-important attribute was thought to be twice as important as the least-important attribute, it would be assigned an initial value of 200; etc.

4. The third-least-important attribute is considered next. Its judged importance also is compared to that of the least-important attribute. If it were thought to be four times as important as the least-important attribute, and the second-least-important attribute were thought to be twice as important as the least-important attribute, the three attributes would be assigned initial importance values of 100, 200, and 400.

5. The judged importance of each attribute of the entity under consideration is determined by following these steps in succession, always working *from* the attribute judged to be least important, *to* the attribute judged to be most important. Attributes judged to be equally important are assigned the same initial importance value.

6. Initial importance weights assigned to the various attributes are then computed by dividing each initial importance value by the sum of the initial importance values.

Edwards and Newman (1982) provide several variants on this ratio weighting procedure. As described above, judges would be instructed to use the least-important attribute as a reference point for each of their subsequent judgments; i.e., judging the importance of the next-least-important attribute in relation to that of the least-important attribute, and then judging the importance of the third-least-important attribute in relation to that of the least-important attribute, and so on. Alternative instructions would have judges use the least-important attribute as a benchmark for the next-least-important, use the next-least-important attribute as a benchmark for the third-least-important, and continue in this fashion until the importance of the most-important attribute had been specified in relation to that of the second-most-important attribute.

Recognizing that the ratio weights that result from these two variants might not be the same, Edwards and Newman suggested a procedure for checking the consistency of importance weights resulting from the ratio method. To assess consistency they had judges repeat the ratio weighting procedure using as benchmarks, the attributes ranked second-least-important, third-least-important, etc. They then compared across the benchmarks, the ratios of weights assigned each pair of attributes. Edwards and Newman suggested that these results be organized in an array they termed a "triangular table."

Policy Capturing Methodology

Hobson and Gibson (1983) have characterized policy capturing as "a general procedure designed to describe statistically the unique information processing strategies of individual raters" (p. 640). In describing the application of policy capturing to the appraisal of personnel, Hobson and Gibson identified the following steps (p. 640):

> (a) the presentation to raters of a series of performance profiles consisting of scores on the major dimensions of performance; (b) instructions to raters to review each profile and then assign an overall rating that best represents or summarizes the available information; and (c) the use of multiple regression analysis to calculate the extent to which an individual rater's overall ratings are predictable, given scores on the separate performance dimensions, and to compute the relative importance of each single dimension in determining overall ratings. The statistical equation resulting from the regression analysis defines the "captured policy" for each individual rater. This captured policy is taken to represent an explicit objective description of the way in which the rater combines and weights dimensional information in arriving at overall ratings.

Although other researchers have introduced variations on the method described by Hobson and Gibson (cf., Barron & John, 1980; Yates & Jagacinski, 1979; Hoffman, 1960; Slovic & Lichtenstein, 1971), their description is essentially

complete. Policy capturing strategies have been used to describe judges' relative weightings of attributes in such areas as clinical judgment (Hoffman, 1960; Hammond, Hursch & Todd, 1964), personnel selection (Anderson, 1977; Naylor & Wherry, 1965; Taylor & Wilsted, 1974; Zedeck & Kafry, 1977), foreign policy beliefs (Summers & Stewart, 1968; Summers, Taliaferro & Fletcher, 1970), factors influencing choice of an organization or position (Feldman & Arnold, 1978; Sheard, 1970; Zedeck, 1977), human learning (Peterson, Hammond & Summers, 1965), factors related to individual motivation (Dachler & Mobley, 1973; Mitchell & Biglan, 1971), and interpersonal conflict (Summers, Stewart & Oncken, 1968). The objectives of policy capturing often include the derivation of one or more "summary" policies that characterize the relative importance weightings of attributes for all raters, or a major category of raters (Bottenberg & Christal, 1968; Zedeck & Kafry, 1977).

Development and Presentation of Policy-Capturing Profiles.

The most popular method used to develop profiles for policy-capturing research is simulation (Naylor & Wherry, 1965; Anderson, 1977; Hobson, Mendel & Gibson, 1981; Stumpf & London, 1981), although Taylor & Wilsted (1974) collected actual profile data in a field setting. When profiles are developed through simulation, the researcher constructs a series of fictitious sets of attribute ratings that are then presented to judges for evaluation and use in rating the entities to which the attributes apply. In most previous studies that used simulated profiles, levels of attributes were selected in such a way that scores on each pair of attributes were uncorrelated across the entire set of profiles. This quest for methodological simplicity, together with the use of simulated profiles, has been criticized by Hobson and Gibson (1983) on the grounds that it limits the generalizability of policy capturing results. These authors also contended that resulting judgments of entities fail to present an accurate paramorphic representation (Hoffman, 1960; Dawes &

Corrigan, 1974) of judges' behaviors when faced with real attribute profiles. Yet Hobson and Gibson concluded, largely on the basis of studies that used simulated profiles of attribute ratings, that (p. 641):

> ... the general linear model has worked well in describing rater policies, as evidenced by the consistently high rater R^2's obtained when regressing overall ratings on scores for the separate dimensions of performance.
> ... there is generally consistent evidence indicating that raters generally possess poor insight into their own rating policies (Hobson, et al, 1981; Taylor & Wilsted, 1974; Zedeck & Kafry, 1977).
> ... in three of four studies utilizing rater clustering procedures (Hobson, et al., 1981; Naylor & Wherry, 1965; Stumpf & London, 1981; Zedeck & Kafry, 1977), researchers found statistically distinct, relatively homogeneous rater subgroups possessing different rating orientations. ... Thus, in general, it has not been possible to use a single rating policy to describe the decision making behavior of all raters.

Arguing in the context of human performance rating, Hobson and Gibson (p. 643) suggested that serious construct validity problems arise when judges are asked to provide summative ratings using simulated profiles of attribute ratings. They contended that hypothetical rating profiles are deficient in terms of (1) volume (i.e., too few descriptors are provided), (2) specificity (simulated profiles describe the quality of attributes too generally), and (3) familiarity (judges asked to provide overall ratings of specific personnel would know those persons individually in a real judgment situation). These objections to the use of simulated profiles are speculative and are not well documented. Counter-arguments in favor of simulated profiles might include: (1) voluminous descriptions are potentially distracting; they might cause judges to consider attributes that are outside the set that should be used in making summative judgments, (2) more-specific descriptions of the quality of attributes might be excessively narrow; they might cause judges to concentrate on specific characteristics or behaviors of the entity to be judged that are

unrepresentative or only partially representative of the attributes to be considered, and (3) familiarity with persons (or entities) to be judged might result in halo effect or the consideration of irrelevant attributes that would undermine the construct validity of the resulting ratings.

Previous studies of policy capturing have used several formats for the presentation of attribute profiles. This variety of approaches and the resulting consistency with which judges adopted rating policies (as indicated by coefficients of determination) are summarized in Table 1. These results suggest that stimulus profiles presented in the form of graphical scales produce more consistent judgments than do verbally-presented profiles.

Methodology

Judgment data were collected from a group of 29 key educational leaders in a southeastern state. Members of the group included the presidents of the state's associations of elementary school principals, middle-school principals, and high school principals, the state's "Teacher of the Year," the president of the state's teachers' association, and the president of the state's association of educational leaders. The group of judges also included other principals and teachers, curriculum specialists, school media specialists, school counselors, and school superintendents.

Multiattribute Utility Technology

In the MAUT study, 29 judges worked independently to rank eight selected components (Reading/Language Arts programs, Mathematics programs, Special Education programs, Guidance and Counseling programs, programs for "At-Risk" students, Personnel Evaluation and Development programs, average Reading Achievement test scores, and average Mathematics Achievement test scores) from "most important" to "least important" in contributing to the overall quality of a school. After assigning a value of 100 to the component he or she judged to be least-

Table 1. Summary of policy-capturing literature in terms of types of judges used, objects of assessment, types of stimulus profiles used, and resulting coefficients of determination (R^2).

Reference	Type of Judge	Object of Assessment	Type of Profile	Coefficient of Determination (R^2)
Zedeck & Kafry (1977)	35 public health nurses	Quality of public health nurses (40 stimuli)	Narrative descriptions	Minimum 0.40 Median 0.61 Maximum 0.71
Summers, Taliaferro, & Fletcher (1970)	131 university under-graduates	175 hypothetical nations -- level of social development	Numerical list	Median 0.56
Stumpf & London (1981)	43 managers	48 hypothetical candidates for promotion to a management position	Narrative description containing numerical ratings	Mean 0.70; zero, after correction for shrinkage
Hobson, Mendel & Gibson (1981)	20 university faculty members	100 hypothetical faculty members rated for performance	Line graph with numerical ratings	Minimum 0.61 Median 0.77; 0.69 when corrected for shrinkage maximum 0.94
Grizzle & Witte (1984)	33 graduate students	80 simulated social service agencies	Visual profiles with numerical ratings	Minimum 0.53 Maximum 0.95
Yates, Jagacinski & Faber (1978)	35 volunteer students	150 hypothetical college courses	Numerical profiles in 5-point Likert scales	Mean 0.81
Anderson (1977)	164 high school teachers	Unspecified	Verbal and numeric (comparison study)	Mean 0.72 (for numeric profiles)

important, each judge assigned values to all other components, working in sequence from the next-least-important, to the most-important. In assigning these values, each judge considered the importance of these components in relation to that of the least-important component. Judges next repeated the process, beginning with the component they judged to be second-least important. Each judge considered the relative importance of the third-least-important component in relation to the second-least-important, and continued in like manner until he or she had judged each successive component in terms of its relative importance, using the second-least-important component as a baseline. Data resulting from this second rating step were used to check the consistency of the judges' initially-assigned importance values. The initially-assigned values were normalized by dividing each initial value by the sum of these values. The second set of values was converted to the scale of the first set and then was normalized. Edwards and Newman (1982) suggested that completion of an entire triangular table might be too laborious if judges were asked to work with more than six or seven attributes. We therefore limited our search for inconsistencies to the first two columns of a triangular table. A sample of our MAUT instruments, together with the written instructions judges received, is available from the first author.

The data produced by this judgment activity were analyzed to determine (1) the relative weights judges assigned to the eight attributes, and (2) the degree of consistency with which judges assigned relative weights to the eight attributes. Normalized weights were computed from judges' assigned importance values in two columns of the triangular table, as described above. For each judge, pairs of normalized weights determined for a given attribute were then averaged. Using these average normalized weights as units of analysis, median normalized weights for each attribute were computed across judges.

Policy Capturing

In our application of policy-capturing methodology, 28 judges were presented with a series of hypothetical profiles that portrayed each of eight school programs and achievement test averages as "non-standard," "standard," or "exemplary." Fifteen judges reviewed profiles for 80 secondary schools and classified these schools as "non-standard," "standard," or "exemplary." Thirteen judges classified 80 elementary schools in the same manner.

On the basis of observations made during a pretest of our instruments, we instructed each judge to begin with profiles for 10 "Preliminary Schools" prior to responding to profiles for the 80 schools used in our analyses. When we observed judges' response latencies during our pretest, it appeared that the judges used an initial set of approximately 10 profiles to establish their policies, and then applied these policies to the remaining profiles. Since our interest focused on policies that were considered and thoughtful, we decided to omit from our analyses, data collected during initial periods of transient policy establishment.

Our examination of policy capturing presented attribute profiles in a graphical format that contained a set of clearly marked, verbally anchored, numerical scales (1=nonstandard; 2=standard; 3=exemplary). This choice of format was based on our review of the literature summarized in Table 1.

Hypothetical profiles were constructed by first generating eight sets of uniformly-distributed random numbers on the closed interval [0,1]. Since the attributes of schools used in this simulation were mutually correlated, it was necessary to construct a set of profiles that reflected this correlational structure. An appropriate correlational structure was supported only in part by relevant research.

Walberg (1982; 1984) reported correlations around 0.35 between average measures of student achievement and various measures of school quality. Jaeger

(1970) reported a correlation of 0.84 between median mathematics achievement test scores and median reading achievement test scores for sixth-graders. Shanahan and Walberg (1985) found that correlations between ratings of school quality (defined as ratings of school-level instruction, the schools' reputations in their communities, and ratings of teachers' interest in students) and average scores on various standardized achievement tests ranged between 0.16 and 0.20 for high school sophomores. Other studies (Walberg & Tsai, 1985; Stahl & Fairbanks, 1986) produced correlations between individual characteristics of school programs (e.g., "Does the school provide reading enrichment classes?") and averaged student achievement test scores that were barely above zero (in the range 0.02 to 0.06). We ignored these latter correlations because individual questions about school programs were so narrowly defined that they did not adequately characterize the quality of an instructional program. In addition, the reliabilities of responses to individual questions are typically low and, as a consequence, attenuate the values of correlations involving these variables.

To construct a matrix of correlations between each pair of attributes, we engaged in a "thought experiment" that incorporated the limited findings of the empirical literature cited above. Based on this partial correlational structure, we postulated the balance of the elements so as to produce a matrix that was positive definite and substantively reasonable.

In the second step of the procedure used to generate hypothetical profiles we conducted a principal components analysis of the postulated matrix of correlations among attributes. The factor structure matrix resulting from this analysis was then postmultiplied by the eight-dimensional matrix of uniform random variates generated in the first step to form an eight-dimensional matrix of random variates with the desired correlational structure (cf., Kaiser & Dickman, 1962).

In the final step of the procedure, the eight-dimensional matrix of random variates resulting from the second step was transformed, so that elements of each vector assumed integer values on the closed interval [1,3].

Multiple linear regression is the analytic procedure of choice throughout the policy-capturing literature. This model assumes that judges regard each rating dimension as contributing linearly and additively to the overall quality of the entities under consideration. The typical justification for the use of an additive multiple linear regression model is the reasonably high coefficients of determination observed in most policy-capturing studies (cf., Zedeck & Kafry, 1977). In addition, investigations of models that incorporate interaction terms have shown extremely small increments in proportions of predicted variation of criterion scores compared to models that are strictly additive (Slovic & Lichenstein, 1971; Slovic, Fischhoff, & Lichtenstein, 1977; Stahl & Zimmerer, 1984). Stahl and Zimmerer justified the elimination of interaction terms from their models by citing the robustness of additive linear models in the behavioral decision theory literature (Beach, 1967; Darlington, 1968; Dawes, 1979; Dawes & Corrigan, 1974; Hoffman, 1960; Laughlin, 1978). They might have supported their argument further by referencing the more than four decades of theoretical work on integration of information in human judgment or decision making (Asch, 1946; Hammond, 1955; Johnson, 1955; Osgood & Tannenbaum, 1955; Peak, 1955; Rimoldi, 1956; Hoffman, 1960; Meehl, 1960; Willis, 1960; Yntema & Torgerson, 1961; Young & Christensen, 1962; Helson, 1964; Spence & Guilford, 1933; Anderson, 1968). Because of the ubiquity of additive linear models in policy-capturing research, it was our model of choice as well.

Once regression coefficients had been estimated for each judge, attribute weights were calculated using a formulation suggested by Hoffman (1960):

$$w_i = \beta_i r_i / R^2$$

where w_i is the weight for the ith attribute, β_i is the standardized regression coefficient for the ith attribute, r_i is the correlation between the dependent variable and the ith attribute (in our models, the dependent variable was the classification of a school as (1) non-standard, (2) standard, or (3) exemplary), and R^2 is the coefficient of determination for the model. We then calculated the median value of the distribution of weights for each attribute and the median value (as well as other descriptive statistics) of the distribution of coefficients of determination, R^2.

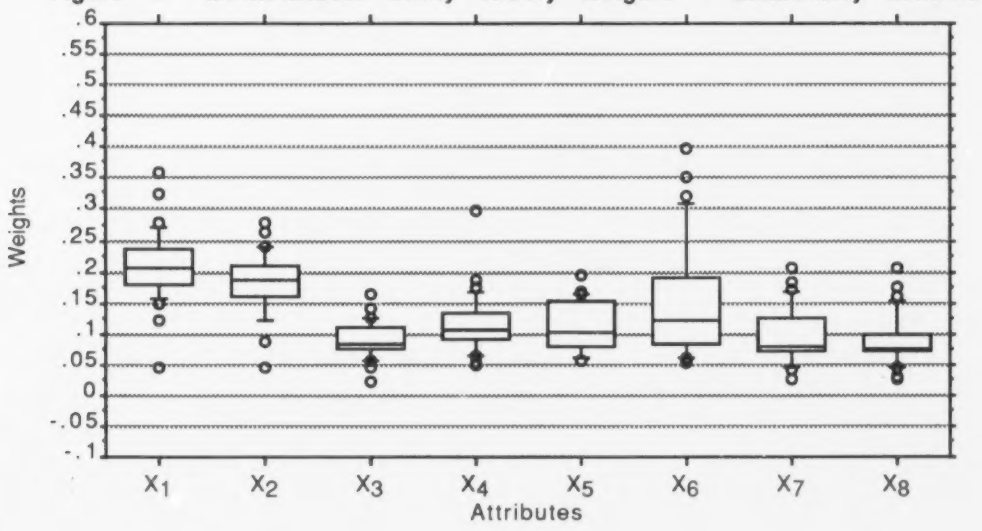
Results

Results derived from application of the MAUT procedure and results derived from application of the policy-capturing procedure are first presented separately. These results are then compared. For each procedure, we present distributions of weights for the eight attributes separately, by level of school. We then compare the distributions of coefficients of determination produced by the policy-capturing procedure with those produced by using the MAUT weights in simulated multiple regression equations.

MAUT Results

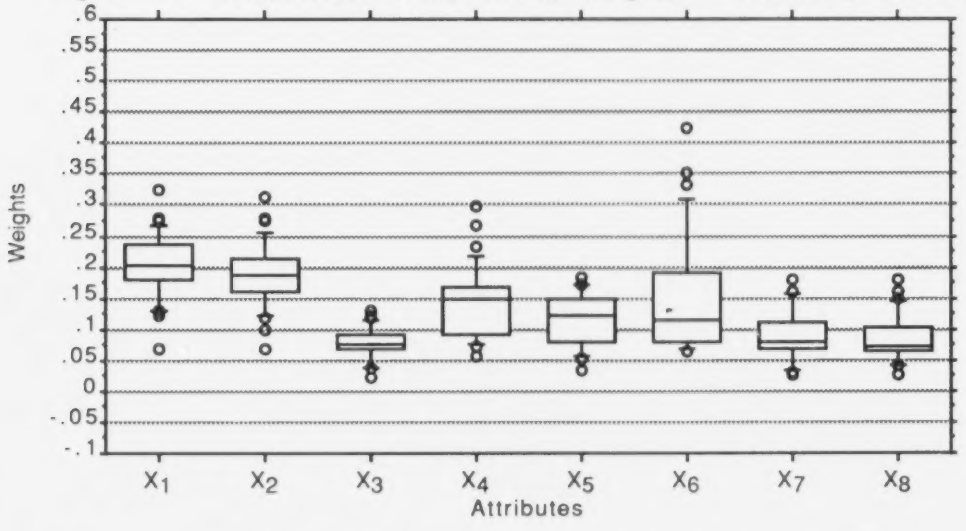
An examination of the median normalized weights assigned by the 29 judges using the MAUT procedure indicates that these weights were not uniform. That is, some attributes of schools were valued far more highly than others when judges considered their relative importance in determining the overall quality of elementary and secondary schools. However, the median weights assigned to specific attributes of elementary schools differed very little from the weights assigned to corresponding attributes of secondary schools; i.e., the median weights for elementary schools (shown in Figure 1) are very similar to those assigned to corresponding attributes of secondary schools (shown in Figure 2).

Figure 1. Multiattribute Utility Theory Weights -- Elementary Schools



Legend: X_1 = Reading Programs; X_2 = Mathematics Programs; X_3 = Guidance and Counseling Programs; X_4 = Special Education Programs; X_5 = Programs for At-Risk Students; X_6 = Personnel Evaluation and Development Programs; X_7 = Students' Average Reading/Language Arts Achievement Test Scores; X_8 = Students' Average Mathematics Achievement Test Scores.

Figure 2. Multiattribute Utility Theory Weights -- Secondary Schools



Legend: X_1 = Reading Programs; X_2 = Mathematics Programs; X_3 = Guidance and Counseling Programs; X_4 = Special Education Programs; X_5 = Programs for At-Risk Students; X_6 = Personnel Evaluation and Development Programs; X_7 = Students' Average Reading/Language Arts Achievement Test Scores; X_8 = Students' Average Mathematics Achievement Test Scores.

At the median, judges' most highly valued attribute (Reading/Language Arts Programs for both elementary and secondary schools) received a weight of 0.20, which was more than twice the size of the weight assigned their least-highly-valued attribute (for elementary schools, average Reading Achievement and Math Achievement received a weight of 0.08; for secondary schools, average Math Achievement received a median weight of 0.07). This finding supports the use of a judgment procedure for determining the relative weights of attributes in determining the overall quality of schools. Uniform weighting of attributes would be erroneous.

It is perhaps not surprising that judges valued the quality of core academic programs (Reading/Language Arts and Mathematics) most highly when considering the overall quality of elementary schools and secondary schools. Perhaps less expected was the relatively high weight judges assigned to the quality of personnel development and evaluation programs as indicators of the quality of elementary schools. Although a substantial body of research has shown that externally-imposed achievement test results are valued far more highly by legislators and members of the public than by school personnel, it was, nonetheless, a revelation to find that judges assigned the lowest weights to students' average reading/language arts achievement test scores and mathematics achievement test scores. This result was consistent for elementary schools and secondary schools.

Judges varied to some degree in the relative weights they assigned the eight attributes of elementary schools. These variations ranged from "modest" to "substantial," with some judges giving far higher weight to some attributes than would be suggested by variations in median values (see Figure 1). By far, the lowest level of consistency was observed for judgments of the relative importance of the quality of personnel evaluation and development programs. Although the median value assigned to this attribute was relatively high (0.13), judges varied substantially

in their assignment of values. Judges were slightly more consistent in the relatively low values they assigned to special education programs, guidance and counseling programs, programs for at-risk students, and students' average mathematics achievement test scores than in the values they assigned the other four attributes.

As shown in Figure 2, our interpretation of the consistency with which judges assigned weights to eight selected attributes of elementary schools applies, with rare exception, to secondary schools as well. Judges were most variable in their assignment of weights to personnel evaluation and development programs. They were somewhat more variable in their assignment of weights to special education programs in secondary schools than in their assignment of weights to special education programs in elementary schools.

Policy Capturing Results

The captured policies of 13 judges who responded to profiles for elementary schools exhibited substantial variation, as shown in the box-and-whisker charts in Figure 3. At the median, judges assigned the highest weights to students' average reading/language arts achievement test scores and average mathematics achievement test scores (the median weight assigned to average reading achievement test scores was 0.20; the median weight assigned to average mathematics achievement test scores was 0.16). The next highest median weights were assigned to the quality of reading programs and the quality of mathematics programs (0.14 for each). Guidance and counseling programs, special education programs, and programs for "at-risk" students were assigned somewhat lower, and very similar median weights (0.08, 0.10 and 0.06, respectively). Personnel evaluation and development programs were assigned the lowest median weight of 0.04.

The lowest levels of consistency in assignment of weights were observed for reading/language arts programs, and reading and mathematics achievement test

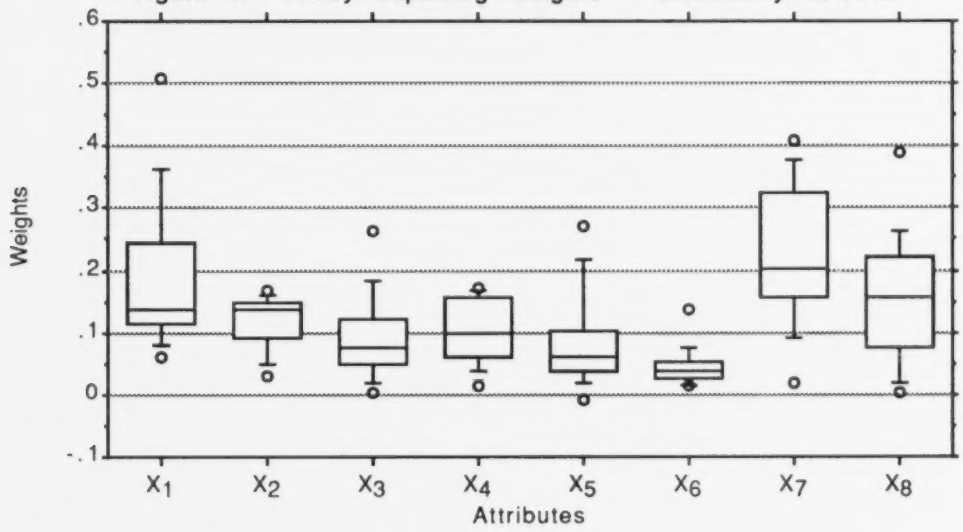
averages. However, the median values assigned to these attributes were relatively high. Judges were most consistent in their weightings of personnel evaluation and development programs; little variation was observed around the low median weight of 0.04.

The captured policies of 15 judges who responded to profiles for secondary schools also exhibited substantial variation, as shown in the box-and-whisker charts in Figure 4. At the median, judges assigned highest weight to the quality of mathematics programs (0.22). Students' average reading/language arts achievement test scores received the next highest median weight (0.20), and the quality of reading/language arts programs received the third-highest median weight (0.18). The other five attributes of secondary schools considered in this study received substantially lower median weights. Students' average mathematics achievement test scores and the quality of programs for at-risk students received respective median weights of 0.11. The quality of guidance and counseling programs, special education programs, and personnel evaluation and development programs were assigned low median weights that were very similar (0.04, 0.04, and 0.05, respectively).

Judges showed substantial inconsistency in their weightings of reading and mathematics programs for secondary schools and in their weightings of reading and mathematics achievement test averages for secondary-school students. However, these attributes were assigned relatively high median weights. Judges were most consistent in their relatively low weightings of guidance and counseling programs.

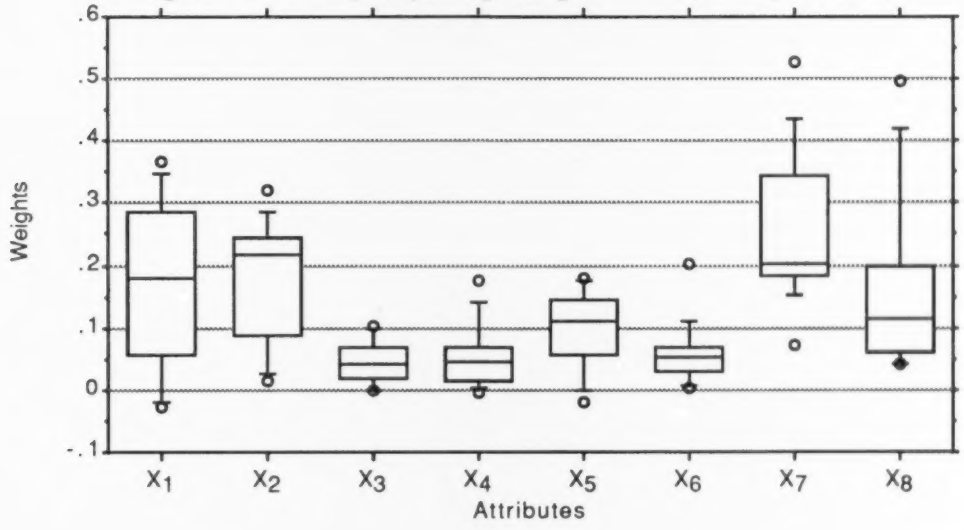
It appears that the distributions of weights assigned to the eight attributes of elementary and secondary schools, as shown in Figures 3 and 4, differ somewhat. However, the samples of judges on which these distributions are based are small, and apparent differences should be interpreted cautiously. Perhaps most notable is the similarity of weights assigned to students' average scores on reading/language

Figure 3. Policy Capturing Weights -- Elementary Schools



Legend: X_1 = Reading Programs; X_2 = Mathematics Programs; X_3 = Guidance and Counseling Programs; X_4 = Special Education Programs; X_5 = Programs for At-Risk Students; X_6 = Personnel Evaluation and Development Programs; X_7 = Students' Average Reading/Language Arts Achievement Test Scores; X_8 = Students' Average Mathematics Achievement Test Scores.

Figure 4. Policy Capturing Weights -- Secondary Schools



Legend: X_1 = Reading Programs; X_2 = Mathematics Programs; X_3 = Guidance and Counseling Programs; X_4 = Special Education Programs; X_5 = Programs for At-Risk Students; X_6 = Personnel Evaluation and Development Programs; X_7 = Students' Average Reading/Language Arts Achievement Test Scores; X_8 = Students' Average Mathematics Achievement Test Scores.

arts achievement tests and mathematics achievement tests when regarded as attributes of elementary schools (medians of 0.20 and 0.16, respectively), and the disparity between the weights assigned to these averages when they were regarded as attributes of secondary schools (medians of 0.20 for average reading achievement and 0.11 for average mathematics achievement). The quality of reading programs and the quality of mathematics programs were assigned relatively high median weights (0.18 and 0.22, respectively) when regarded as attributes of secondary schools, but these attributes were assigned substantially lower median weights when regarded as characteristics of elementary schools (respective weights of 0.14). Guidance and counseling programs and special education programs received higher median weights when regarded as attributes of elementary schools; in contrast, programs for at-risk students received a higher median weight when regarded as an attribute of secondary schools. Personnel evaluation and development programs received similar median weights when regarded as characteristics of elementary schools or secondary schools. The median weights assigned to the quality of personnel evaluation and development programs were lowest among those judges assigned the eight attributes of elementary and secondary schools.

Intrajudge Consistency: MAUT and Policy Capturing

Coefficients of Determination for Policy Capturing. As noted earlier, the coefficient of determination associated with a judge's captured policy can be interpreted as an indicator of the judge's consistency. The coefficients of determination associated with the policies of the 13 judges who responded to profiles for elementary schools ranged from 0.53 to 0.91 with a mean of 0.73. The coefficients of determination associated with the policies of the 15 judges who responded to profiles for secondary schools ranged from 0.35 to 0.87 with a mean of 0.69 (see Table 2). In the distribution of coefficients for judges who responded to

Table 2. Minimum, mean, and maximum coefficients of determination (R^2) resulting from policy capturing and multiattribute utility theory methods of weighting eight components of school quality.

Method	Minimum	Mean	Maximum
Policy Capturing (Elementary Schools)	.53	.73	.91
Multiattribute Utility Theory (Elementary Schools)	.40	.64	.84
Difference (Policy Capturing Minus Multiattribute Utility Theory)	.04	.08	.13
Policy Capturing (Secondary Schools)	.35	.69	.87
Multiattribute Utility Theory (Secondary Schools)	.28	.55	.76
Difference (Policy Capturing Minus Multiattribute Utility Theory)	.02	.15	.54

profiles for secondary schools, the minimum value was an outlier; the next lowest value in the distribution was 0.48.

The coefficients of determination found in these analyses compare favorably with those found in the eight policy-capturing studies described earlier. Recall that the range of coefficients reported by Zedeck and Kafry (1977) was 0.40 to 0.71; Hobson, et. al (1981) reported a range of 0.61 to 0.94; and Grizzle and Witte (1984) reported a range of 0.53 to 0.95. The distribution of median coefficients of determination reported in the eight studies ranged from 0.56 to 0.81, and the median of the distribution of medians was 0.72, a value that is just below the smallest of the two medians (0. 77 and 0.74 for elementary school profiles and secondary school profiles, respectively) found in our analyses.

Coefficients of Determination for MAUT. Coefficients of determination were also computed by using the MAUT weights provided by judges who had completed both MAUT and policy-capturing judgment exercises. The MAUT weights provided by each of these judges (n=25) were applied to the attribute ratings of each policy-capturing profile to compute a predicted summative rating of each hypothetical elementary or secondary school. These predicted ratings were then correlated with judges' actual policy-capturing ratings of the same schools. Coefficients of determination (R^2) were computed by squaring these correlations.

Coefficients of determination based on weights produced by 12 judges who applied the MAUT procedure to attributes of elementary schools ranged from 0.40 to 0.84, with a mean of 0.64. Coefficients based on weights produced by 13 judges who applied the MAUT procedure to attributes of secondary schools ranged from 0.28 to 0.76, with a mean of 0.55. These results are shown in Table 2.

Comparison of Coefficients of Determination Across Methods. Table 2 also contains minimum, mean, and maximum differences between coefficients of

determination derived from the policy capturing and MAUT procedures. Results are shown separately for elementary schools and secondary schools.

The difference between the coefficients of determination produced by the policy-capturing and MAUT methods was computed for each judge. For elementary schools, this difference ranged from a minimum of 0.04 to a maximum of 0.13, with a mean of 0.08. This mean difference was statistically significant ($t = 7.819$, $\alpha = 0.05$). For secondary schools, the difference between the coefficients of determination produced by the policy-capturing and MAUT methods ranged from a minimum of 0.02 to a maximum of 0.54, with a mean of 0.15. The maximum difference of 0.54 was an outlier; the next largest difference was 0.37. The mean R^2 difference of 0.15 for secondary schools was also statistically significant ($t = 3.3123$, $\alpha = 0.05$).

Discussion

Weights obtained from both the MAUT procedure and the policy-capturing procedure provide evidence that assignment of equal weights to the eight attributes of elementary and secondary schools examined here would misrepresent the values judges ascribed to these attributes as contributors to the overall quality of schools. Each procedure produced distributions of median weights that varied substantially from a uniform distribution.

The MAUT procedure and the policy-capturing procedure produced distributions of median weights that differed markedly. The inconsistencies found in previous comparative studies of the MAUT and policy-capturing procedures (Hobson, et al., 1981; Taylor & Wilsted, 1974; Zedeck & Kafry, 1977) were echoed in the findings of the current study. Most noticeably, judges who participated in the MAUT exercises assigned relatively low weights to students' average performances on reading/language arts achievement tests and mathematics achievement tests (median weights of 0.08 were assigned to each as attributes of elementary schools;

median weights of 0.08 and 0.07 were assigned to these attributes as characteristics of secondary schools). In contrast, as a result of the policy-capturing exercises, students' average performances on reading/language arts achievement tests and mathematics achievement tests were assigned median weights of 0.18 for elementary schools, and median weights of 0.20 and 0.11 for secondary schools. The distributions of weights judges assigned attributes of schools using multiattribute utility technology were not consistent with the distributions evidenced in their captured policies. The holistic judgments required by the policy-capturing procedure evidently evoked different responses than did the pairwise-comparative judgments required by MAUT.

As indicators of judges' rating consistencies, coefficients of determination (R^2) that resulted from application of the policy-capturing procedure were moderate to high -- a mean of 0.73 for elementary schools and a mean of 0.69 for secondary schools. The R^2's that resulted from use of the MAUT procedure were substantially lower -- a mean of 0.64 for elementary schools and a mean of 0.55 for secondary schools. The differences between these means were statistically significant for both elementary and secondary schools. If consistency of judgment were used as a criterion for selection of an attribute weighting method, policy capturing would be the method of choice. This conclusion is in keeping with the findings of studies cited earlier; demonstration of its generalization to judgments of the attributes of elementary and secondary schools is, we believe, unprecedented.

REFERENCES

Anderson, B. L. (1977). Differences in teachers' judgment policies for varying numbers of verbal and numerical clues. *Organizational Behavior and Human Performance, 29,* 68-88.

Anderson, N. H. (1968). A simple model for information integration. In R.P. Abelson, E. Aronson, W. J. McGuire, T. M. Newcomb, M. J. Rosenberg & P. H. Tannenbaum (Eds.) *Theories of cognitive consistency: A sourcebook.* Chicago: Rand McNally, 731-743.

Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology, 41,* 258-290.

Barron, F. H. & John, R. (1980). Reference effects: A sheep in wolf's clothing. *Organizational Behavior and Human Performance, 25,* 365-374.

Beach, L. R. (1967). Multiple regression as a mo del for human information utilization. *Organizational Behavior and Human Performance, 2,* 276-289.

Bottenberg, R. A. & Christal, R. E. (1968). Grouping criteria -- a method which retains maximum predictive efficiency. *Journal of Experimental Education, 36,* 28-34.

Cattin, P. (1980). Estimation of the predictive power of a regression model. *Journal of Applied Psychology, 65,* 407-414.

Dachler, H. P. & Mobley, W. H. (1973). Construct validation of an instrumentality-expectancy-task-goal model of work motivation: Some theoretical boundary conditions. *Journal of Applied Psychology, 58,* 397-418.

Darlington, R. B. (1968). Multiple regression in psychological research and practice. *Psychological Bulletin, 69,* 161-182.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34,* 571-582.

Dawes, R. M. & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81,* 95-106.

Dyer, J. S. & Sarin, R. K. (1979). Measurable multiattribute value functions. *Operations Research, 27,* 810-822.

Edwards, W. (1977). How to use multiattribute utility measurement for social decision-making. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-7,* 326-340.

Edwards, W. & Newman, J. R. (1982). *Multiattribute evaluation.* Beverly Hills, CA: Sage Publications.

Eils, L. C. & John, R. S. (1980). A criterion validation of multiattribute utility analysis and of group communication strategy. *Organizational Behavior and Human Performance, 25,* 268-288.

Einhorn, H. J. & McCoach, W. (1977) A simple multiattribute utility procedure for evaluation. *Behavioral Science, 22,* 270-282.

Feldman, D. C. & Arnold, H. J. (1978). Position choice: Comparing the importance of organizational and job factors. *Journal of Applied Psychology, 63,* 706-710.

Gardiner, P. C. & Edwards, W. (1975). Public values: Multiattribute utility measurement for social decision making. In M. F. Kaplan & S. Schwartz (Eds.). *Human judgment and decision processes.* New York: Academic Press.

Grizzle, G. A. & Witte, A. D. (1984). Evaluating multidimensional performance: A social judgment theory approach. *Evaluation Review, 8,* 777-800.

Hammond, K. R. (1955). Probabilistic functioning and the clinical method. *Psychological Review, 62,* 255-262.

Hammond, K. R., Hursch, C. & Todd, F. (1964). Analyzing the components of clinical inference. *Psychological Review, 71,* 438-456.

Helson, H. (1964). *Adaptation-level theory.* New York: Harper and Row.

Hobson, C. J., Mendel, R. M. & Gibson, F. W. (1981). Clarifying performance appraisal criteria. *Organizational Behavior and Human Performance, 28,* 164-188.

Hobson, C. J. & Gibson, F. W. (1983). Policy capturing as an approach to understanding and improving performance appraisal: A review of the literature. *Academy of Management Review, 8,* 640-649.

Hoffman, P. J. (1960). The paramorphic representation of clinical judgment. *Psychological Bulletin, 47,* 116-131.

Jaeger, R. M. (1984). *Sampling in education and the social sciences.* New York: Longman.

Johnson, D. M. (1955). *The psychology of thought and judgment.* New York: Harper and Row.

Kaiser, H. F. & Dickman, K. (1962). Sample and population score matrices and sample correlation matrices from an arbitrary population correlation matrix. *Psychometrika, 27,* 179-182.

Keeney, R. L. (1977). The art of assessing multiattribute utility functions. *Organizational Behavior and Human Performance, 19,* 267-310.

Keeney, R. L. (1982). Decision analysis: An overview. *Operations Research, 30,* 803-838.

Laughlin, J. E. (1978). Comment on "Estimating coefficients in linear models: It don't make no nevermind." *Psychological Bulletin, 85,* 247-253.

Meehl, P. E. (1960). The cognitive activity of the clinician. *American Psychologist, 15,* 19-27.

Mitchell, T. R. & Biglan, A. H. (1971). Instrumentality theories: Current uses in psychology. *Psychological Bulletin, 76,* 432-454.

Naylor, J. C. & Wherry, R. J. (1965). The use of simulated stimuli and the "JAN" technique to capture and cluster the policies of raters. *Educational and Psychological Measurement, 25,* 969-986.

Osgood, C. E., & Tannenbaum, P. H. (1955). The principle of congruity in the prediction of attitude change. *Psychological Review, 62,* 42-55.

Peak, H. (1955). Attitude and motivation. In M. Jones (Ed.), *Nebraska symposium of motivation, 1955.* Lincoln, Nebraska: University of Nebraska Press, 149-188.

Pedhazur, E. J. (1982). *Multiple regression in behavioral research* (2nd. ed.). New York: Holt, Rinehart and Winston.

Peterson, C. R., Hammond, K. R., & Summers, D. A. (1965). Optimal cue weighting in multiple probability learning. *Journal of Experimental Psychology, 70,* 270-276.

Pitz, G. F. (1980). Assessing the utility of multiattribute utility assessments. *Organizational Behavior and Human Performance, 26,* 65-80.

Rimoldi, H. J. A. (1956). Prediction of scale values for combined stimuli. *British Journal of Statistical Psychology, 9,* 29-40.

Shanahan, T. & Walberg, H. J. (1985). Productive influences on high school student achievement. *Journal of Educational Research, 78,* 357-363.

Sheard, J. L. (1970). Intrasubject prediction of preferences for organizational types. *Journal of Applied Psychology, 54,* 248-252.

Slovic, P., Fischhoff, B., & Lichtenstein, S. (1977). Behavioral decision theory. *Annual Review of Psychology, 28,* 1-39.

Slovic, P. & Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance, 6,* 649-744.

Spence, W. & Guilford, J. P. (1933). The affective value of combinations of odors. *American Journal of Psychology, 45,* 495-501.

Stahl, S. A. & Fairbanks, M. M. (1986). The effects of vocabulary instruction: A model-based meta-analysis. *Review of Educational Research, 56,* 72-110.

Stahl, M. J. & Zimmerer, T. W. (1984). Modeling strategic acquisition policies: A simulation of executives' acquisition decisions. *Academy of Management Journal, 27,* 369-383.

Stumpf, S. A. & London, M. (1981). Capturing rater policies in evaluating candidates for promotion. *Academy of Management Journal, 24,* 752-766.

Summers, D. A. & Stewart, T. R. (1968). Regression models for foreign policy beliefs. *Proceedings of the 76th Annual Convention of the American Psychological Association,* 195-196.

Summers, D. A., Stewart, T. R., & Oncken, G. R. (1968). Interpersonal conflict in heterocultural dyads. *International Journal of Psychology, 3,* 191-196.

Summers, D. A., Taliaferro, J. D., & Fletcher, D. J. (1970). Subjective vs. objective description of judgment policy. *Psychonomic Science, 18,* 249-250.

Taylor, R. L. & Wilsted, W. D. (1974). Capturing judgment policies: A field study of performance appraisal. *Academy of Management Journal, 17,* 440-449.

Walberg, H. J. & Tsai, S. (1985). Correlates of reading achievement and attitude: A National Assessment study. *Journal of Educational Research, 78,* 159-167.

Wiener, R. L., Frank, M. A. & Libros, T. (1988). Termination guidelines for a magnet school review committee: Application of decision theory. *Professional Psychology: Research and Practice, 19,* 316-322.

Willis, R. H. (1960). Stimulus pooling and social perception. *Journal of Abnormal and Social Psychology, 60,* 365-373.

Yates, J. F. & Jagacinski, C. M. (1979). Reference effects in multiattribute evaluations. *Organizational Behavior and Human Performance, 24,* 400-410.

Yates, J. F., Jagacinski, C. M., & Faber, M. D. (1978). Evaluation of partially described multiattribute options. *Organizational Behavior and Human Performance, 21,* 240-251.

Yntema, D. B., & Torgerson, W. S. (1961). Man-computer cooperation in decisions requiring common sense. *IRE Transactions of the Professional Group on Human Factors in Electronics, 2,* 20-26.

Young, P. T. & Christensen, K. R. (1962). Algebraic summation of hedonic processes. *Journal of Comparative and Physiological Psychology, 55,* 332-336.

Zedeck, S. (1977). An information processing model and approach in the study of motivation. *Organizational Behavior and Human Performance, 18,* 47-77.

Zedeck, S. & Kafry, D. (1977). Capturing rater policies for processing evaluation data. *Organizational Behavior and Human Performance, 18,* 269-294.

Handout for:

Alternative Procedures for Integrating Multidimensional Evaluations of Schools: An Experimental Comparison

Richard M. Jaeger
Claire H. Usher
Center for Educational Research and Evaluation
University of North Carolina at Greensboro

Presented at the 1991 Annual Meeting of the American Educational Research Association, Chicago, April 3-7

Determining the Relative Importance of Attributes using MAUT.

Several procedures have been proposed by Edwards and Newman (1982) for determining the relative importance of attributes to an overall judgment of the quality of an entity. The one used for this study, which involves independent consideration of the relative importance of any pair of attributes, is termed "ratio weighting," and consists of the following steps:

1. The attributes are listed in rank order of their judged importance, with the least-important attribute listed first.

2. The least-important attribute is assigned an initial importance value of 100.

3. The next-least-important attribute is assigned an initial value that reflects its relative importance, compared to the least-important attribute. Thus, if the next-least-important attribute was thought to be half again as important as the least-important attribute, it would be assigned an initial value of 150, which is 50 percent larger than the value assigned to the least-important attribute (100). If the next-least-important attribute was thought to be twice as important as the least-important attribute, it would be assigned an initial value of 200; etc.

4. The third-least-important attribute is considered next. Its judged importance also is compared to that of the least-important attribute. If it were thought to be four times as important as the least-important attribute, and the second-least-important attribute were thought to be twice as important as the least-important attribute, the three attributes would be assigned initial importance values of 100, 200, and 400.

5. The judged importance of each attribute of the entity under consideration is determined by following these steps in succession, always working *from* the attribute judged to be least important, *to* the attribute judged to be most important. Attributes judged to be equally important are assigned the same initial importance value.

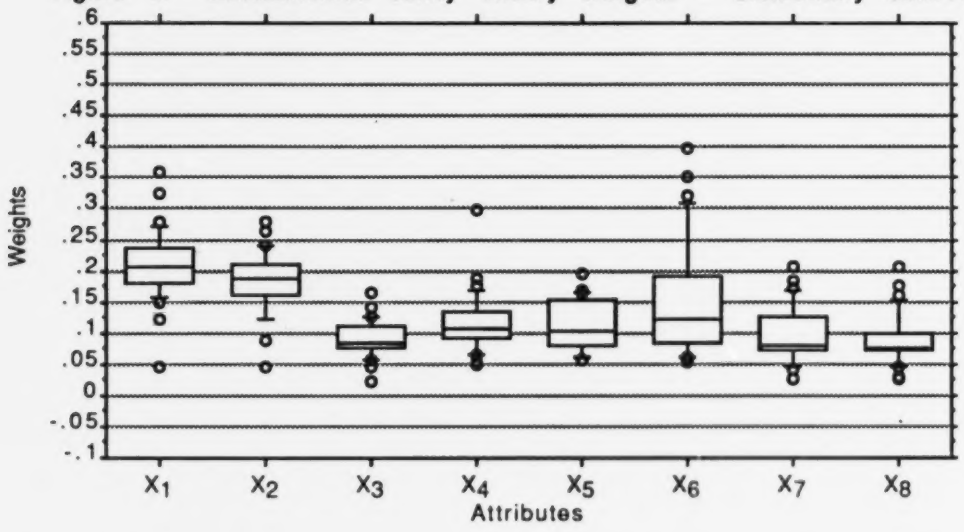
6. Initial importance weights assigned to the various attributes are then computed by dividing each initial importance value by the sum of the initial importance values.

Policy Capturing Methodology

Hobson and Gibson (1983) have characterized policy capturing as "a general procedure designed to describe statistically the unique information processing strategies of individual raters" (p. 640). In describing the application of policy capturing to the appraisal of personnel, Hobson and Gibson identified the following steps (p. 640):

> (a) the presentation to raters of a series of performance profiles consisting of scores on the major dimensions of performance; (b) instructions to raters to review each profile and then assign an overall rating that best represents or summarizes the available information; and (c) the use of multiple regression analysis to calculate the extent to which an individual rater's overall ratings are predictable, given scores on the separate performance dimensions, and to compute the relative importance of each single dimension in determining overall ratings. The statistical equation resulting from the regression analysis defines the "captured policy" for each individual rater. This captured policy is taken to represent an explicit objective description of the way in which the rater combines and weights dimensional information in arriving at overall ratings.

Figure 1. Multiattribute Utility Theory Weights -- Elementary Schools



Legend: X_1 = Reading Programs; X_2 = Mathematics Programs; X_3 = Guidance and Counseling Programs; X_4 = Special Education Programs; X_5 = Programs for At-Risk Students; X_6 = Personnel Evaluation and Development Programs; X_7 = Students' Average Reading/Language Arts Achievement Test Scores; X_8 = Students' Average Mathematics Achievement Test Scores.

Figure 2. Multiattribute Utility Theory Weights -- Secondary Schools

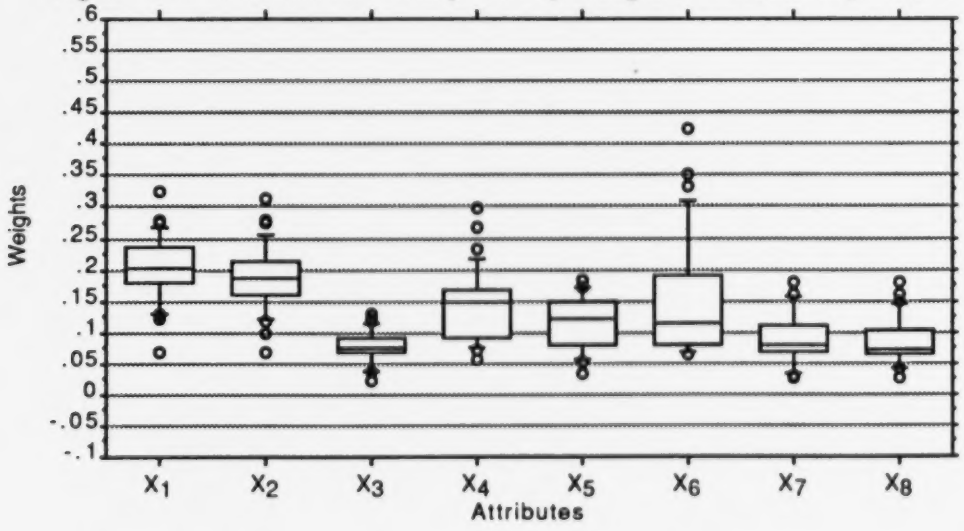
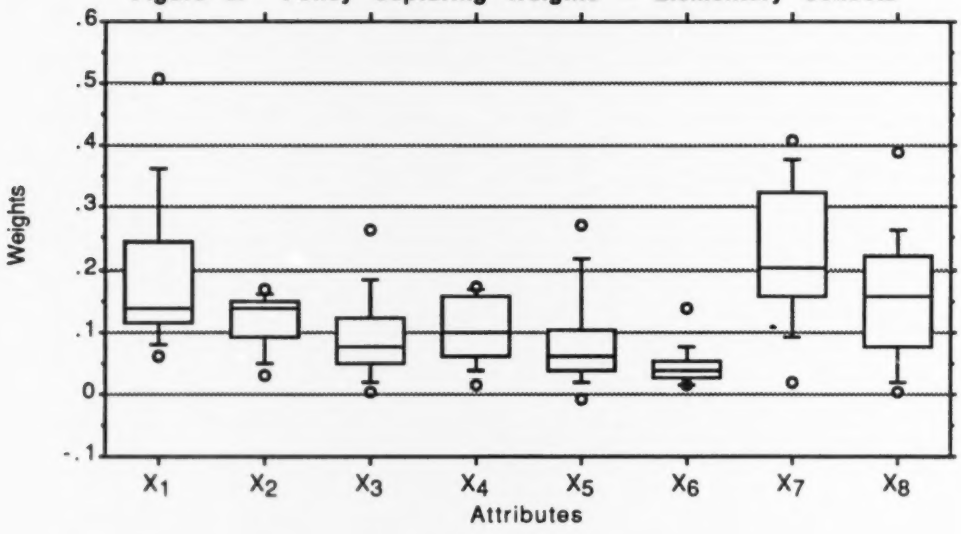


Figure 3. Policy Capturing Weights -- Elementary Schools



Legend: X_1 = Reading Programs; X_2 = Mathematics Programs; X_3 = Guidance and Counseling Programs; X_4 = Special Education Programs; X_5 = Programs for At-Risk Students; X_6 = Personnel Evaluation and Development Programs; X_7 = Students' Average Reading/Language Arts Achievement Test Scores; X_8 = Students' Average Mathematics Achievement Test Scores.

Figure 4. Policy Capturing Weights -- Secondary Schools

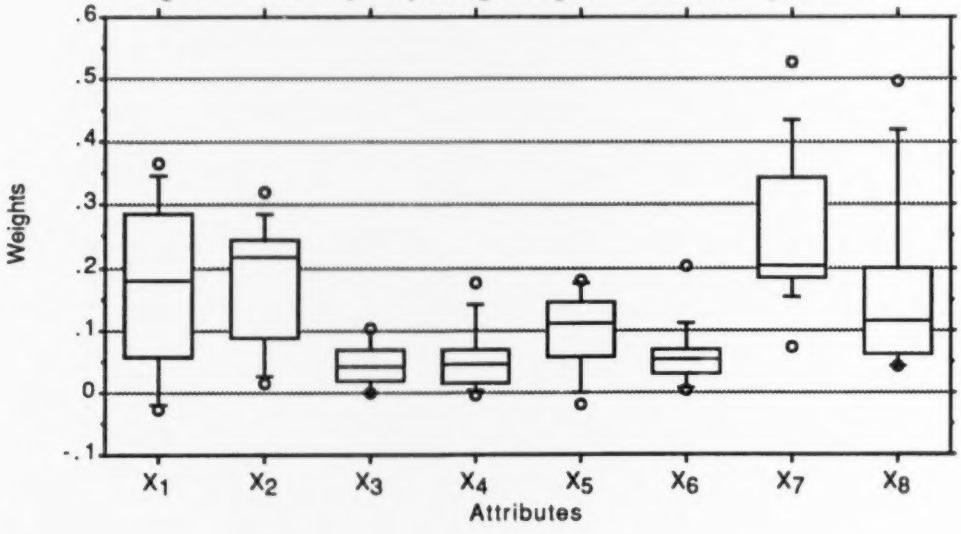


Table 2. Minimum, mean, and maximum coefficients of determination (R^2) resulting from policy capturing and multiattribute utility theory methods of weighting eight components of school quality.

Method	Minimum	Mean	Maximum
Policy Capturing (Elementary Schools)	.53	.73	.91
Multiattribute Utility Theory (Elementary Schools)	.40	.64	.84
Difference (Policy Capturing Minus Multiattribute Utility Theory)	.04	.08	.13
Policy Capturing (Secondary Schools)	.35	.69	.87
Multiattribute Utility Theory (Secondary Schools)	.28	.55	.76
Difference (Policy Capturing Minus Multiattribute Utility Theory)	.02	.15	.54